International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Data Preparation and Analysis for Andhra Pradesh Clusters

Srinatha Karur[a]*, Prof. M.V. Ramana Murthy[b]

[a]*Oracle DBA & IT faculty, Government Engineering College, Ibra, P0. 327,Sultanate of Oman*

[b]*Professor and Director, Department of Computer Science, Osmania University, Hyderabad, India*

[a]*Email:karur_sri@yahoo.co.in*

[b]*Email: mv_rm@rediffmail.com*

Abstract

Local clusters are highly preferable in all domains due to complex and large Database applications are available. Clustering techniques are applied to local clusters as per needs of local clusters. We can apply divide and conquer rule for local clusters. Local clusters are always constructed as per needs of local bodies. In future we can combine or integrate these local clusters with big clusters or centralized clusters. The number of local cluster formation is completely depend upon requirements of local bodies. But in some contexts they must work along with central systems when they are integrated or combine with central systems.

Keywords:Regression Analysis; Curve fitting; Data mining tools; Clustering techniques;

1. Introduction

1.1 Geographical description of Andhra Pradesh

Andhra Pradesh is a one of the important states in the Republic of India where multi language, culture, religion, geographical conditions, Universities etc.is available. Due to its complete heterogeneous nature and logical conditions in India no single policy should work perfectly.So different policies and methods are necessary to implement the strategic planning.The below figure shows the geographical map of Andhra Pradesh (A.P) which is a part of India as one of the State.

---

* Corresponding author.
E-mail address: karur_sri@yahoo.co.in.

Fig. 1.Shows different districts of A.PState

1.2 Zonal description of Andhra Pradesh

Even though there are 23 Units are available in A.P technically only 20 nodes are available. This is due to zero sample is available in remaining Districts (Units).The authors already discussed about the Indian Universities in their publication [1]. The State capital city is Hyderabad and generally it is treated as root node of all nodes which are available in a state. But our policy is mainly construction of local clusters and examines the data with respect to local policy or policies but not with central policy. So generally the node Hyderabad is considered as a maximum occurring node when compare to other nodes due to its high priory nature. The author divided the whole unit as six local units. This initial split occurred on the basis of continuous occurrences .Nearest neighborhood is treated as strategy for split the unit. This strategy may or may work out for large data sampling. The author already discussed about the different levels of University clustering in their publication [1], and way of sampling for India country level [2].The districts (local clusters) which are available in A.P state as shown in the figure and also from A.P official website [20].

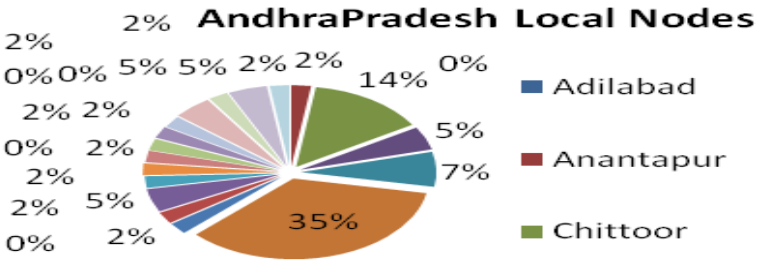


Fig. 2. The figure shows Pie chart of different districts with number of Universities

Kurnool, Anantapur, Cuddapah and Chittor are called as “Rayalaseema” unit. Hyderabad, Warangal, Mahaboob Nagar, Nalgonda Khamam, Karimnagar, Medak, Adilabad and Nizamabad are called as “Telangaana” unit. Now Hyderabad unit is spited as “RangaReddy and Hyderabad” units. So total local units for “Telangana” have ten only.Nellore, Prakasam, Guntur, Krishna, West Godavari, East Godavari, Vishakhapatnam, Vijayanagaram and Srikakulam are termed as “Coastal Andhra”. So broadly there are three 3 local clusters are available in Andhra Pradesh [13].Once again 10 districts of “Costa Andhra” region we can form two groups. One is coastal area nodes and others are called “Andhra”. Six nodes belong to coastal region and four nodes belongs to “Andhra” (Non-coastal area).

From “Rayalaseema” region all nodes formed local clusters and from “Coastal Andhra” region “Vijayanagarm “and Prakasam” nodes have zero sample and from “Telangana” region “Medak, Khamam and Adilabad” have zero samples. In these districts University is not established. The below table gives the exact idea of NULL sampling nodes from “Andhra Pradesh” state.

Table-1.Shows Null nodes in A.P State, India.

Sno	Region	NULL nodes	Short Name
1	Rayalaseema	0(>0)	NA
2	CoastaAndhra	2	VN,Vizag
3	Telangana	3	MDK,AD,KHM

Sometimes it is necessary or context come for local clusters are merged with external clusters and that situation is called “logical clusters” in the view of local clusters. Conversion of geographical clusters into logical clusters is out of scope of this context. Within local clusters we can form different clusters with different domains. For example University-Industry clusters within defined local unit is called local hybrid clusters. The authors already discussed different types of Innovative clusters in their publication [11].

We can use Regression analysis and Data mining tools for Mathematical and technical implementation of local clusters formation in the view of Andhra Pradesh Universities. The Mathematical models are implemented with different technologies for good quality and maximum Graphical User Interface. We combined here both Mathematical modeling and Data mining tools for best accuracy and understand of local clusters in the view of Andhra Pradesh Universities.

2. Material and Methodology

2.1 Resources and Methods for Mathematical Modeling

“Regression Analysis” and “Curve fitting” methods are very important methods for estimate the relation between two or more variables in a given or applied sample. We can found” Regression Analysis” in various Statistical and Mathematical applications. Generally “Curve fitting” method belongs to domain of “Numerical methods”, which deals about the accuracy of solution in terms of digits itself only. Practically in the clustering policy we cannot implement this amount of accuracy especially in the field of Telecommunications domain where Networking is lion share of concept. In terms of technical it is necessary to store the data in “Excel sheets” and applied the method explicit or implicit [3].

Curve fittings the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, possibly subject to constraints. Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a "smooth" function is constructed that necessary to analyze or study the relation between two or more variables. This study turns to Regression Analysis for estimate the relation and meaning between two or more variables which are connected by some Mathematical formula. In statistics, regression analysis

is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quintile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution [5].

The implementation of 'n' variables relation is called n Th degree polynomial is out of scope from this context. Due to technology changes and strong Mathematics background of authors giving more and more user GUI's.Lot of .NET oriented tools are available for "Regression Analysis and Curve fitting methods". Now the tools are available for both Non-linear and linear "Regression Analysis". Now more predictions and confidence is available with .NET oriented tools [7].

Some websites are giving the services on different domains with respect to Mathematics and Statistics flavor. Especially they are working for Data Analysis for big scale applications and giving service for execution of data [8].

Some websites are extremely providing online tutorials on complete Mathematical models with practical examples. They can provide all academic audio and video for registered users. They have their own software for implement the Mathematical and Statistical Models as per their company policies. Origin Lab is very nice software which is available as educational and commercial versions [9].

In real life sometimes it is necessary to analyze about non-linear curves also. Generally non-linear Regression curves are constructed with Successive approximation method which is the out of scope of context. But we cannot neglect the concept of non-linear. Since in real life applications are linear or non-linear or hybrid. In all domains we have linear and nonlinear applications. For example in Education domain time table planning is linear whereas administrative matters are non-linear. As per Data structures concepts non-linear nature has tree structure and linear are available side by side in neighborhood fashion. In Engineering applications generally we want 3D nature also. Generally dealing with 3 variables have very lengthy manual process. Separate software are available for 3D linear and non-linear regression methods. NLREG is a powerful statistical analysis program that performs linear and nonlinear regression analysis, surface and curve fitting. NLREG determines the values of parameters for an equation, whose form you specify, that cause the equation to best fit a set of data values. NLREG can handle linear, polynomial, exponential, logistic, periodic, and general nonlinear functions. Unlike many "nonlinear" regression programs that can only handle a limited set of function forms, NLREG can handle essentially any function whose form you can specify algebraically. NLREG also computes auxiliary statistics. The Standard version of NLREG can fit up to 5 variables and parameters to the data observations. The Advanced version can handle up to 2000 variables and parameters. In addition, the advanced version can generate 3D surface plots such as shown here. In this case X,Y and Z axis are available[11][22]. The details of implementation are out of scope of context.

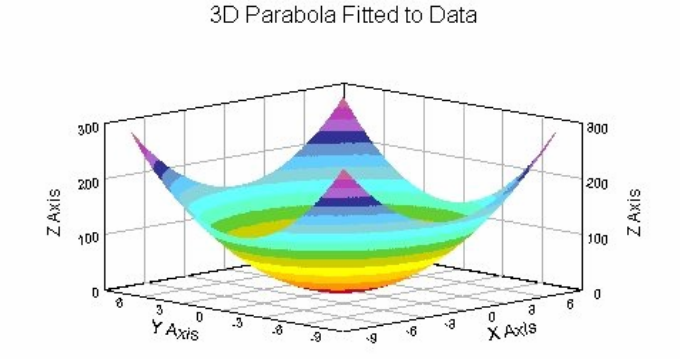


Fig 3. 3D nature of given problem.

Some of the tools are available with high GUI with user interface and implicitly combined with other applications. Sigma plot tool is very nice GUI with user interact nature. All results are copied into another application such as MS-Office, and other user applications. We can integrate software with MS-Office in all respects for provide high degree of GUI and user interact [12].

Some web sites are available in such a way that they are allow users to define own macros and use Excel sheets very effectively. They are allowing only Excel sheets for data storing. With Visual Basic script we can define our own macros for given and defined problems. They are not allowing Active-X controls. More details are available in [13]. The goal of linear regression is to adjust the values of slope and intercept to find the line that best predicts Y from X. More precisely, the goal of regression is to minimize thesum of the squares of the vertical distances of the points from the line. The Graph pad is GUI version for estimate the linear regression and curve fitting methods for given data[14].Some tools are providing customization with inbuilt models. Approximately 90 models are available for better of customer. This tool is called “Curve Expert” and allowed users with all maximum combinations. This tool is designed on the basis of automatic best fit method [15].

Now days in a market lot of data mining tools are available. Out of them some are freeware and some are shareware. Some of data mining tools are especially launched by companies such as Microsoft data mining tools which comes along with MS-SQL server. For research and Education they are providing free service with limited data option. It is available in Excel –add in along with MS-SQL server any version. We can download SQL-Server data mining add-in from Microsoft official website [16].The tools which are available with Microsoft is Server oriented which means that client should have server version also which is generally not required.

Lot of free data mining tools are available in the market with very minimum service charges and free of cost also. Depend upon nature of data we may take service from site owners. For that we pay some reasonable price. Tanagra, Orange, Weka, R, and Rapid miner are the very good and most popular free wares for data mining. The website[17] shows combination of free wares and shareware software.

2.2 Resources and tools for Data mining techniques

Data mining techniques are available as supervised and unsupervised learning. Separate statistical methods are necessary to model supervised learning. All Data mining tools are used Statistical methods for supervised learning. Tanagra is very nice Data mining tools which is highly preferred by Research community. Maximum all supervised learning methods are implemented in pure Statistics methods. More details are available from website[18].This is available in both French and English versions. But this software is not flexible and sometimes we don’t know the reasons for errors generation. For each and every time we must take fresh copy and then implement the tool. This repetitive work creates some inconvenience to user. This project is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering,

factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. The last modification of site by author is, April 23, 2008.In Tanagra the Data is processing on the basis of "Continuous and Discrete" which leads confusion to user. Statistics certificate course helps user to understand exact nature of input requirements. In Tanagra different types of clustering methods are available as shown in table-2.But in this software we have Principal Component Analysis as "Factorial Analysis". The main formats of input file for "Tanagra" are *.txt,*,xls and *.arff. Tanagra allows very strictly *.xls but not higher versions (only 2003 but not 2007 and 2010).We can load the database into "Tanagra" is as follows.

	Nodes	Hdyerabad	Kurnool	Anatapur	Chittor	Cuddapah
1	Hdyerabad	0	250	400	575	415
2	Kurnool	250	0	175	360	200
3	Anatapur	400	175	0	250	160
4	Chittor	575	360	245	0	165
5	Cuddapah	425	200	155	165	0

Fig 4. Shows data is loaded into Tanagra tool

The problems in Tanagra tool can be easily handled by all advanced Data mining Java flavor tools. They are executing on different types of data. If Data is very big and different fields Tanagra is not suitable and we can switch over to other Data mining tools such as Orange, Weka, and Rapid Miner etc. Orange providing very powerful canvas called Orange Canvas and script is available in "Python". Especially it gives very beautiful "Visualization" effects. Before and after the process we can apply these "Visualization" effects for easy understanding of user. Authors applied "Hierarchical Clustering" with Orange tool only. We can measure the quality of classification methods are by using "Test learners" only. For more details on Orange visit official website[20]. The authors are observed that, with all combinations of methods in Hierarchical clustering the result is same as shown in the figure 3. This implementation is in Orange tool.

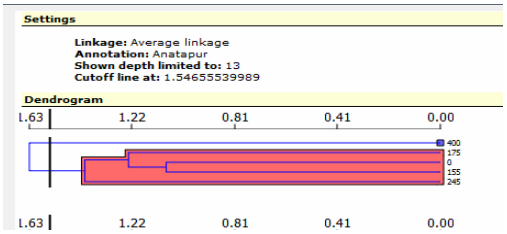


Fig 5. Shows the depth is limited to 13 for all.

The core Java tool is Weka which has very good GUI and it has different tools such as "SQLViewer, package manager, Bayes Network, and arff converter". The allowed data put is *.txt,*.csv,*.arff form only for data analysis.*.arff is the standard Weka type file which is useful for all tools almost. Detail description of file formats is out of scope. This software is available under GNU public license. In Weka we can use either open source code or use tool for data analysis. In Weka the important tools are package manager, arff viewer, SQL viewer and Bayes net viewer. Visualization tools are also available as shown in Fig.

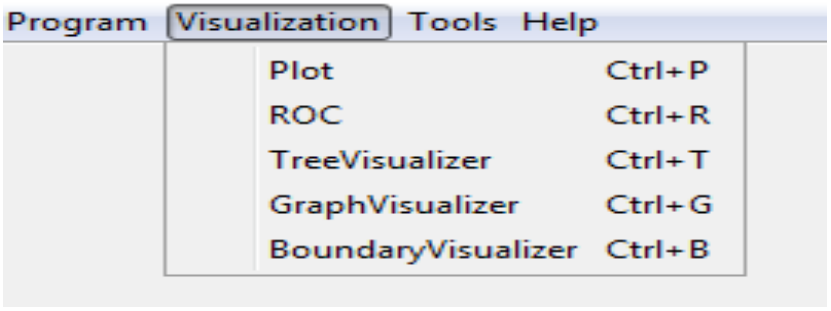


Fig 6. Shows different visualization tools are in Weka

For more details visit Weka's official website [19].Unlike Orange Weka is more user-friendly. We can directly get the results using GUI mode of Weka. Mediators have no role in Weka. The important format of input data files are *.arff,*.csv and *.txt (tab delimited). In Weka command line also available. The below table shows the "Rayalaseema" local nodes are loaded into Weka with Weight is1. In "Rayalaseema" four nodes are available as shown in the below Fig. The node "Hyderabad" is treated as default node" and generally itis common to all areas of Andhra Pradesh due to attribute values is as "CapitalCity" of Andhra Pradesh.

No.	Label	Count	Weight
1	Hdyerabad	1	1.0
2	Kurnool	1	1.0
3	Anatapur	1	1.0
4	Chittor	1	1.0
5	Cuddapah	1	1.0
6		1	1.0

Fig 7. Shows database is loaded into Weka tool

Rapid Miner is a complete business analytics workbench with a strong focus on data mining, text mining, and predictive analytics. It uses a wide variety of descriptive and predictive techniques to give you the insight to make profitable decisions. Rapid Miner together with its analytical server Rapid Analytics also offers full reporting and dash boarding capabilities and, therefore, a complete business Intelligence solution in combination with predictive analytics. The important input data file format are *.txt(tab delimited and *.arff). The authors observed that the proposed system data model is not supported by Rapid miner as shown in the Fig.

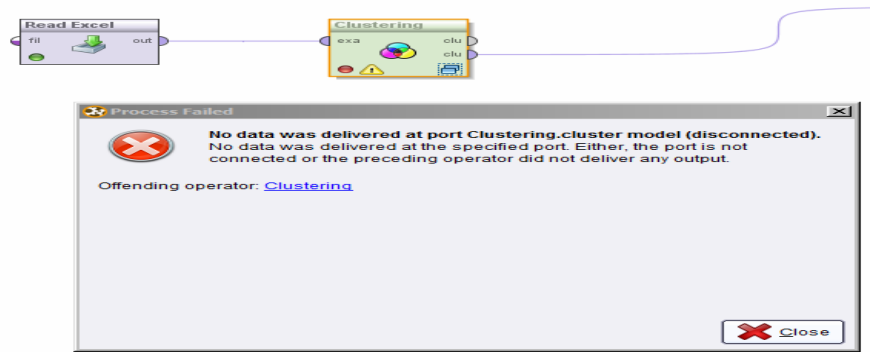


Fig8. Shows the proposed dataset refused by Rapid miner.

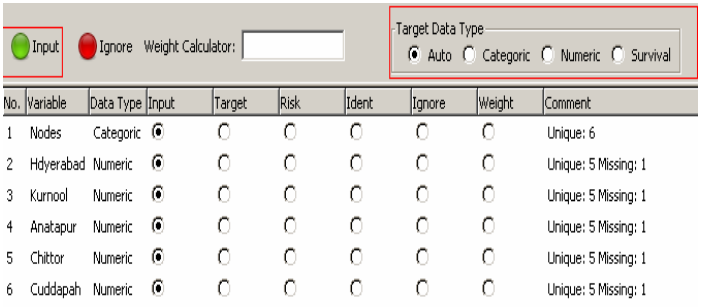
R is also very good Data mining tool which is useful for even Fuzzy clustering also. As per our need we can download dynamically the packages from R website. Lot of mirrors are available for downloading. Basically “R” is command line tool which is little bit tricky and hard also. But we can use package called “Rattle” download directly and convert R as GUI. We can use these method for download the packages as follows.

>install.packages(“rattle”)

>library(“rattle”)

>rattle()

If all things are holds good we can get R package GUI called “Rattle” GUI for communicate with R asa end user. We can use all facilities which are available in R tool. For more details refer R official journal.



No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Nodes	Categoric	●	○	○	○	○	○	Unique: 6
2	Hdyerabad	Numeric	●	○	○	○	○	○	Unique: 5 Missing: 1
3	Kurnool	Numeric	●	○	○	○	○	○	Unique: 5 Missing: 1
4	Anatapur	Numeric	●	○	○	○	○	○	Unique: 5 Missing: 1
5	Chittor	Numeric	●	○	○	○	○	○	Unique: 5 Missing: 1
6	Cuddapah	Numeric	●	○	○	○	○	○	Unique: 5 Missing: 1

Fig 9. Data is loaded into R.

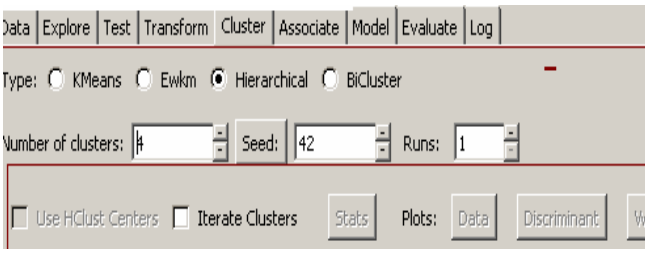


Fig 10. Shows database is loaded but not work.

3. Results and Reports

3.1 Graphical representation of Regression Analysis for Local Clusters

The regression analysis of the entire state is conducted on the basis of the data available in UGC official websites. The number of Universities which are available in respect location is treated as a node. In some Districts no Universities are available. They are called zero nodes. In Rayalaseema only all n>0, and only four districts are available. In another two areas called "Coastal Andhra"& "Telangana" zero nodes are available. The following curves show Regression Analysis of A.P State, India. The below table shows various linear and polynomial equations and R^2 values of various local clusters within Andhra Pradesh, India. In the below figure both linear and polynomial curves are drawn with R^2 values. Both linear and polynomial equations are obtained and for polynomial negative values are generated due to x2< 0. So "Rayalaseema" area for a polynomial is neglected and linear curve values are >0 so it is valid as shown in figure-11.From below figure it is observed that for polynomial this data is not suitable.

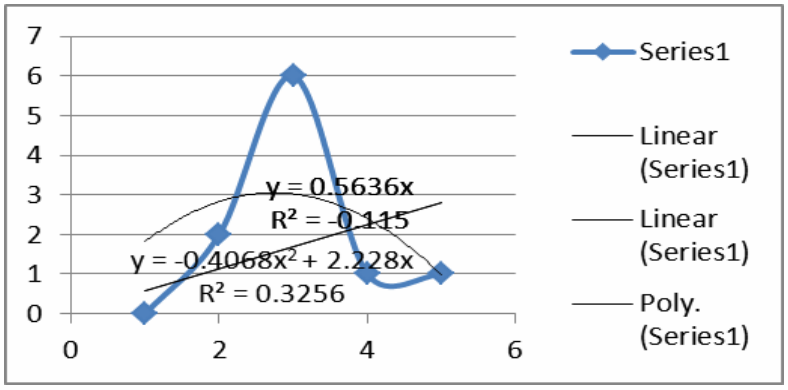


Fig 11:Shows Regression and Curve fitting for Rayalaseema area as single cluster with 4 local clusters

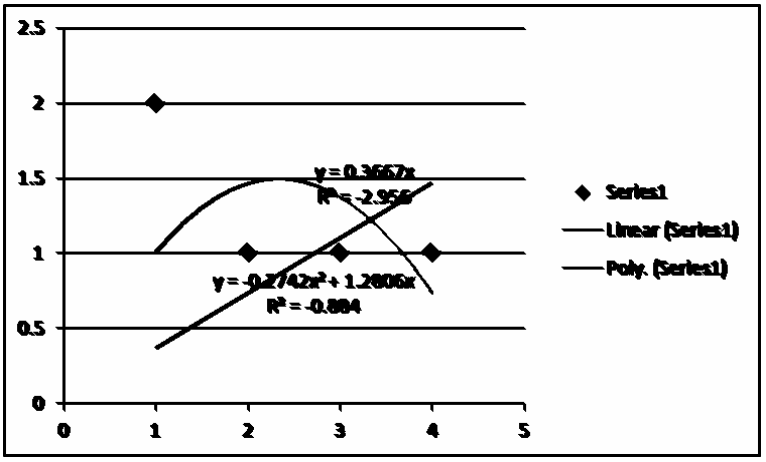


Fig 12.Regression and curve fitting for Coastal Andhra.

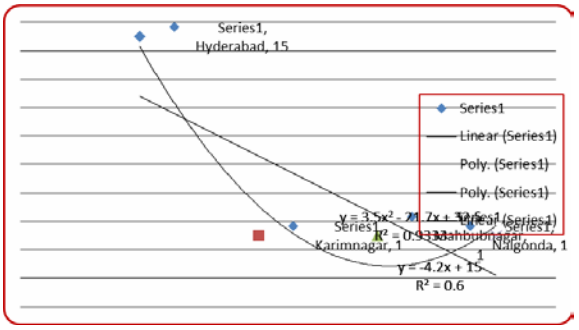


Fig 13.Regression and curve fitting for Andhra region.

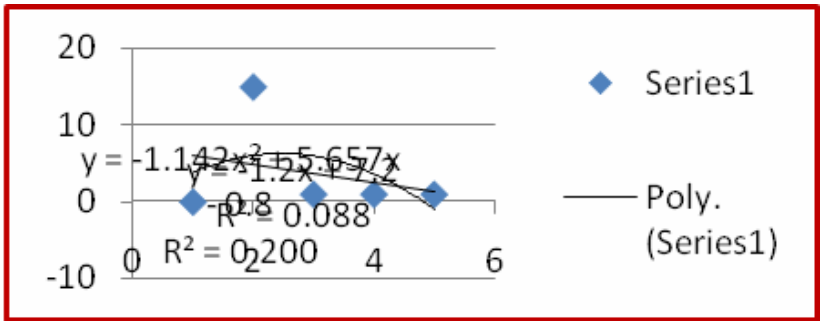


Fig 14.Regression and Curve fitting for Telangana-1 area.

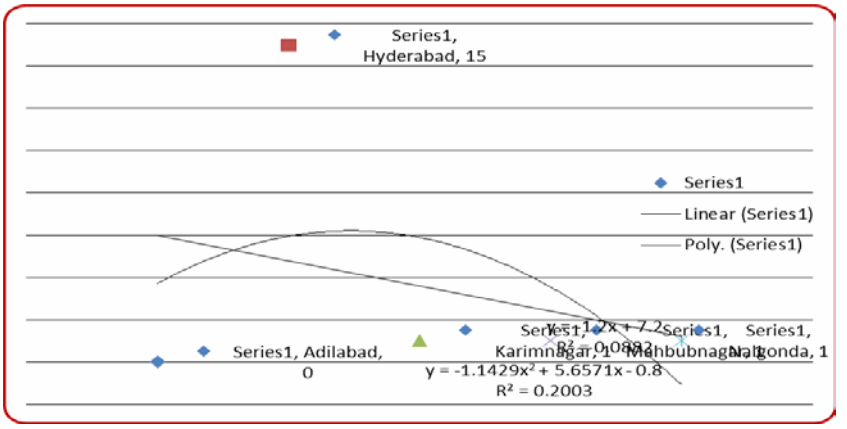


Fig15.Regression and Curve fitting for Telangana-2 area.

The following table-2 shows Report of Regression Analysis and Curve fitting of local clusters.(Linear fitting)

Table 2.The below table shows the Regression Analysis for A.P Clusters (Linear and Polynomial)

Sno	Local Cluster Name	Linear		Polynomial	
		y	R2	Y	R2
1	Rayalaseema	0.5636x	0.115	$-0.41x^2+2x$	0.32356
2	Andhra	0.3667x	-2.956	$-0.2142x^2+1.2606x$	0.884
3	Coastal Andhra	-2x+2	0.3333	NA	NA
4	Telangana-1	-1.2x+7.2	0.0882	$-1.143x^2+5.671x-0.8$	0.2003
5	Telangana-2	0.5x+4.3333	0.0036	$-14.5x^2-8.7x-14$	1

The local clusters of “TABLE-3” area called “Rayalaseema” region clusters have the following cut-off values as shown in figure 3.The following table shows the result of remaining local clusters with in “Rayalaseema” area.

Table 3.Rayalaseema Clusters

City	Single	Complete	Average	Wards
Kurnool	1.3856	1.711	1.546	1.794
Cuddapah	1.3856	1.711	1.546	1.794
Anatapur	1.3856	1.711	1.546	1.794
Chittor	1.3856	1.711	1.546	1.794

Table 4.Andhra Clusters

City	Single	Complete	Average	Wards
Nellore	1.4473	1.5435	1.4954	1.6431
Guntur	1.4473	1.5435	1.4954	1.6431
Machalipatnam	1.4473	1.5435	1.4954	1.6431

Table.5. Coastal Andhra Clusters

City	Single	Complete	Average	Wards
Eluru	1.3472	1.6956	1.5454	1.7603
Kakinada	1.3472	1.6956	1.5454	1.7603
Vizag	1.3472	1.6956	1.5454	1.7603
V.Nagaram	1.3472	1.6956	1.5454	1.7603
Srikakulam	1.3472	1.6956	1.5454	1.7603

Table 6.Telangana-1 Clusters

City	Single	Complete	Average	Wards
Hyderabad	1.3114	1.6887	1.4866	1.8219
M.Nagar	1.3114	1.6887	1.4866	1.8219
K.Nagar	1.3114	1.6887	1.4866	1.8219
Nalgonda	1.3114	1.6887	1.4866	1.8219

In Telangana-2, onlyNizamabad has >0 entity. Remaining members have zero entity. So by default it creates Bi-cluster with Hyderabad local cluster.

Table 7.Result of Andhra Pradesh Local Clusters for Universities.

Sno	Cluster Name	Depth	Cut-off
1	Rayalaseema	13	1.5456
2	Andhra	13	1.5323
3	Coastal A.P	14	1.5871
4	Telangana-1	14	1.5772
5	Telangana-2	14	1.3435

Acknowledgements

I would like to say thanks to all my teachers from school level to Research level and including non-teaching staff also. I would like to extend my thanks to C.M.JUniversity management to gave this opportunity to do Research in their organization.C.M.JUniversity is located in Shillong, Himachal Pradesh, India. Special thanks to Mr.PraveenRaagi, Publications Expert ,Govt Engineering College,Ibra, Sultanate of Oman, for his support in final document preparation.

Conclusion

From the TABLE-6 it is observed that Rayalaseema and Andhra local clusters can form Bi-clusters and Coastal A.P and Telangana-1 can form Bi-clusters.Telangana-2 become single cluster with default cluster Hyderabad.Telangana-2 has default and implicit cluster otherwise it becomes idle cluster. All linear and polynomial values are shown in Graph. The equation of Telangana-2 $-14.5x^2-8.7x-14$(Figure 15)or TABLE-2 says that mathematically Telenagana-1 and Telangana-2 clusters (for polynomial) should not form since $x^2<0$,is meaningless. So we can form two Bi-clusters as shown in Table-7 on the basis of linear relation (Mathematically) and from tools. In terms of Regression Analysis only linear curve holds good and polynomial curve has $x^2<0$ which makes contradiction.

References

[1] Mr.Srinatha Karur, Prof. M.V.Ramana Muthy,"Survey and Analysis of University Clustering." International Journal of Artificial Intelligence And Applications, vol. 4, pp. 128-144, July. 2013.
[2] Mr.Srinatha Karur,Pof. M.V.Ramana Murthy,"Creation of local Clusters for Indian Universities", International Journal of Artificial Intellegence and Applications", vol. 4, pp. 19-38, September-2013.
[3] Regression Analysis. " Curve fitting tool Box," USENET: http://www.mathworks.com/products/datasheets/pdf/curve-fitting-toolbox.pdf
[4] Curve fitting Methods, " Curve fitting methods," USENET: http://www.curvefitting.com/download.htm
[5] Regression Analysis, "Regression Analysis," USENET: http://en.wikipedia.org/wiki/Regression_analysis
[6] Curve fitting Methods, " Curve fitting Methods," USENET: http://en.wikipedia.org/wiki/Curve_fitting
[7] Curve fitting " Curve fitting Classes," USENET: http://www.extremeoptimization.com/solutions/CurveFitting.aspx?gclid=COX_tK7Rz7kCFYWz3godKgoA2w
[8] USENET: http://alison.com/courses/Diploma-in-Statistics?gclid=CNyhu-nRz7kCFYZa3godFz0Arw
[9] Curve fitting methods, "Curve fitting Methods Using Originallab tool," USENET : http://www.originlab.com/index.aspx?go=Support/VideoTutorials&ss=chm&pid=1564
[10] Non linear regression methods, " Non Linear Regression Methods," USENET: http://en.wikipedia.org/wiki/Nonlinear_regression
[11] Curve fitting Methods, " Curve fitting Methods," USENET: http://www.sigmaplot.com/products/sigmaplot/curvefitting.php
[12] Curve fitting and Regression Analysis,"Model data using for Curve fitting and Regression Analysis," USENET: http://www.mathworks.com/help/exlink/model-data-sets-using-regression-and-curve-fitting.html
[13] Curve fitting, " Curve fitting Principles," USENET: http://www.graphpad.com/guides/prism/6/curve-fitting/
[14] http://www.curveexpert.net/
[15] Ricco Rakotomalala," Linear Regression Analysis," USENET: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[16] MS-SQL Datamining addin," Supervised Methods," USENET: http://www.microsoft.com
[17] Shareware and Freeware datamining tools, " Free datamining tools," USENET: http://www.kdnuggets.com/fdsearch/search.pl?Match=1&Realm=Software_etc&Terms=free
[18] Orange Regression Analysis, " Regression Methods and Applications," USENET: http://www.youtube.com/watch?v=LxGylXe7RMk
[19] Tools for Visulaizations,"Weka tools," USENET: http://www.cs.waikato.ac.nz
[20] USENET: http://www.aponline.gov.in/quick%20links/APFactFile/info%20on%20districts/Districts.htm